Coding of Tree Source (Algebraic Aspects of Coding Theory and Cryptography)

AUTHOR(S):

Kobayashi, Kingo; Morita, Hiroyoshi; Hoshi, Mamoru

数理解析研究所講究録 1361 巻 2004 年 70-79

木情報源の符号化

Coding of Tree Source

電気通信大学・情報通信工学科　小林 欣吾 (Kingo Kobayashi) *
森田 啓義 (Hiroyoshi Morita), 星 守 (Mamoru Hoshi)
University of Electro-Communications

概 要

We introduce a model of (i.i.d.) tree source with respect to a distribution P on non-negative integers and discuss on its entropy. Furthermore, we give the expression of the probability of obtaining infinite tree, that is, of penetrating to infinity without termination for any probability distribution P.

1 Preliminaries

In the study of computer science and information theory, there are many occasions when we encounter combinatorial structures called trees. Most common trees appearing in this field are the rooted ordered trees. We simply denote them as trees in this paper. It would be quite important to devise efficient mechanisms to encode them for many applications such as data compression.

When we studied the pre-order coding of binary tree, we found an interesting identity [1] with respect to Catalan numbers, that is:

Theorem 1

$$\sum_{n=0}^{\infty} \frac{1}{2n+1} \binom{2n+1}{n} 2^{-(2n+1)} = 1. \tag{1}$$

The following proof provides the speed of convergence of summation to the limit one.

*東京都調布市調布ヶ丘 1-5-1. kingo@ice.uec.ac.jp

Proof Let $a_n = c_{2,n}4^{-n}$, where $c_{2,n} = \frac{1}{2n+1}\binom{2n+1}{n}$ is the Catalan number. Then we find that a_n satisfies

$$\begin{cases} (2n+4)a_{n+1} &= (2n+1)a_n, n \geq 0 \\ a_0 &= 1. \end{cases} \tag{2}$$

Moreover, letting $b_n = (2n+1)a_n$, we have the recurrence

$$b_{n+1} + a_{n+1} = b_n \qquad \text{for } n \geq 0 \text{ with } b_0 = a_0 = 1. \tag{3}$$

By summing up (3) from $n = 0$ to N, we obtain

$$b_N + \sum_{n=1}^{N} a_n = b_0.$$

Therefore,

$$\begin{aligned} \sum_{n=0}^{N} a_n &= a_0 + b_0 - b_N \\ &= 2 - \binom{2N+1}{N} 4^{-N}. \end{aligned} \tag{4}$$

From Stirling's formula, the second term of (4) can be expressed by

$$\frac{2}{\sqrt{\pi N^3}}\left(1 + O(1/N)\right). \tag{5}$$

Since (5) goes to zero as $N \to \infty$, the theorem holds. □

This identity means that the pre-order coding for binary trees shows the best possible performance in the sense that its length function tightly satisfies the Kraft inequality.

On the other hand, we have shown inequalities [1] for cases of $k \geq 3$:

$$\frac{1}{2} < \sum_{n=0}^{\infty} c_{k,n} 2^{-(kn+1)} < 1, \tag{6}$$

where the $c_{k,n}$'s are the generalized Catalan numbers (see the definition (8) in the next section). The above inequalities guarantee the existence of a prefix code with the length function $kn+1$ for k-ary trees with n internal nodes, but unfortunately denies that of a code with the length function kn. With respect this point, refer to the remark 1.

The aim of this paper is to show that the identity (1) can be generalized as in the next equation:

$$\sum_{n=0}^{\infty} \frac{1}{2n+1} \binom{2n+1}{n} p^n q^{n+1} = \begin{cases} 1 & \text{for } 0 \le p \le 1/2 \\ \frac{q}{p} & \text{for } 1/2 \le p \le 1 \end{cases}, \tag{7}$$

where $q = 1 - p$. Thus, the case $p = 1/2$ of the identity (1) corresponds to the critical point separating the conditions in the equation (7).

2 A Model of Tree Source

Assume that a probability distribution $P = (p_0, p_1, \ldots)$ on the set of non-negative integers is given. Starting from the root, let us extend each node by s branches and produce s children with probability p_s. Thus, the node will be a leaf (terminal node) with probability p_0. Here, we independently throw a dice to determine the number of extending branches for each surviving node with the identical distribution P. Thus, we define a source model of generating trees by independently using an identical distribution P at each node. The distribution P is called the branching distribution. In this paper, we study the condition on P for eventually getting a finite tree with probability one, and the entropy of tree. Moreover, we provide the probability of getting an infinite tree.

3 Percolation Model on k-ary Tree

Let us restrict the tree source to a special case, that is, a stochastic generation of a k-ary tree. Here, we denote a k-ary tree to be a rooted ordered tree, each internal node of which has k distinct branches, usually corresponding to k characters in an alphabet. Starting from the root, extend k branches and append k children with probability p, or terminate with probability $q = 1 - p$. Then, we have two distinct events. One is the event E_f that we ultimately obtain a finite tree, and the other one is the event E_∞ that the coin flipping process will never be stopped, and we have an infinite tree.

From the argument by Raney, the number $c_{k,n}$ of k-ary tree having n internal nodes is given by

$$c_{k,n} = \frac{1}{kn+1} \binom{kn+1}{n}. \tag{8}$$

Using the generalized Catalan numbers, we can express the probability of the event E_f as

$$\Pr\{\mathrm{E}_f\} = \sum_{n=0}^{\infty} c_{k,n} p^n q^{(k-1)n+1}. \tag{9}$$

In order to evaluate the series of the above equation, let us introduce the generating function $F_{k,p}(z)$ as follows.

$$F_{k,p}(z) = \sum_{n=0}^{\infty} c_{k,n} p^n q^{(k-1)n+1} z^n. \tag{10}$$

Thus,

$$\Pr\{\mathrm{E}_f\} = F_{k,p}(1). \tag{11}$$

With respect to this generating function, we can easily find the functional equation by the symbolic consideration.

$$F_{k,p}(z) = q + pz F_{k,p}(z)^k. \tag{12}$$

For the case $k = 2$, we can explicitly solve the functional equation as follows.

$$\begin{aligned} F_{2,p}(1) &= \frac{1 - \sqrt{1-4pq}}{2p} \\ &= \frac{1 - |2p-1|}{2p} \\ &= \begin{cases} 1 & \text{for } 0 \le p \le 1/2 \\ \frac{q}{p} & \text{for } 1/2 \le p \le 1 \end{cases}. \end{aligned} \tag{13}$$

Also, for the case $k = 3$,

$$\begin{aligned} F_{3,p}(1) &= \frac{\sqrt{p^2+4pq} - p}{2p} \\ &= \begin{cases} 1 & \text{for } 0 \le p \le 1/3 \\ \frac{\sqrt{4p-3p^2}-p}{2p} & \text{for } 1/3 \le p \le 1 \end{cases}. \end{aligned} \tag{14}$$

In general, we have

Theorem 2 The probability of the event E_f of having a finite k-ary tree for the extending probability p is given by

$$\begin{aligned} \Pr\{\mathrm{E}_f\} &= F_{k,p}(1) \\ &= \begin{cases} 1 & \text{for } 0 \le p \le 1/k \\ f(p) & \text{for } 1/k \le p \le 1 \end{cases}, \end{aligned} \tag{15}$$

where $f(p)$ is a unique real value f in the interval [0,1] satisfying the equation,

$$f^{k-1} + f^{k-2} + \cdots + f + 1 = \frac{1}{p}, \tag{16}$$

for $1/k \leq p \leq 1$.

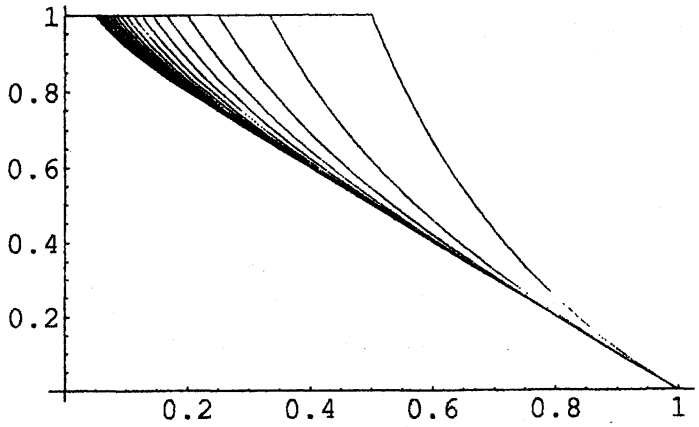

Fig.1 Probability of getting a finite k-ary tree
versus the extending probability p,
(the curves correspond to the cases of $k = 2, 3, \ldots$ from the right).

Remark 1 Previously, we showed an identity [2] with respect to the generalized Catalan numbers,

$$\sum_{n=0}^{\infty} c_{k,n} 2^{-\{g(k)n + \log_2(k/(k-1))\}} = 1, \tag{17}$$

where $g(k) = k\log_2 k - (k-1)\log_2(k-1) = kh(1/k)$ and $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function. The LHS of the above equation is rewritten as

$$\sum_{n=0}^{\infty} c_{k,n} (\frac{1}{k})^n (\frac{k-1}{k})^{(k-1)n+1} = 1. \tag{18}$$

Thus, the identity (17) corresponds to the critical case $p = 1/k$ of the equation (15).

4 Ideal Codeword Length for k-ary Tree

For case $0 \leq p \leq 1/k$, we will eventually have a finite k-ary tree with probability 1. At that time, we can consider that the tree with n internal

nodes has been produced with the probability $p^n q^{(k-1)n+1}$. Here, we notice that the number of leaves (or external nodes) is $(k-1)n+1$. Thus, the ideal length of a codeword for representing the k-ary tree is $-(\log p+(k-1)\log q)n-\log q$. The expectation $\overline{L}$ of the ideal codeword length is given by

$$\overline{L}=\sum_{n=0}^{\infty} c_{k,n}p^n q^{(k-1)n+1}\{-(\log p+(k-1)\log q)n-\log q\}. \tag{19}$$

This expectation should be considered to be the *entropy* of a tree generated in our percolation model.

Therefore, we have to evaluate the sum,

$$\sum_{n=0}^{\infty} c_{k,n}p^n q^{(k-1)n+1}n=F'_{p,k}(1). \tag{20}$$

Differentiating the functional equation (12), we get

$$F'_{p,k}(1)=\frac{p}{1-kp}, \tag{21}$$

for the case $0\le p\le 1/k$. Inserting this evaluation into the equation (19), we obtain

$$\begin{aligned}\overline{L} &= -(\log p+(k-1)\log q)\frac{p}{1-kp}-\log q \\ &= \frac{h(p)}{1-kp}.\end{aligned} \tag{22}$$

The variance $\mathrm{var}(L)$ is calculated by

$$\begin{aligned}&\mathrm{var}(L)\\ &= \sum_{n=0}^{\infty} c_{k,n}p^n q^{(k-1)n+1}\{-(\log p+(k-1)\log q)n-\log q-\overline{L}\}^2\\ &= \sum_{n=0}^{\infty} c_{k,n}p^n q^{(k-1)n+1}\{(\log p+(k-1)\log q)^2n^2\\ &\qquad +2\log q(\log p+(k-1)\log q)n+(\log q)^2\}-\overline{L}^2.\end{aligned} \tag{23}$$

Here, we notice from the functional equation (12) that

$$\sum_{n=0}^{\infty} c_{k,n}p^n q^{(k-1)n+1}n^2=F'_{k,p}(1)+F''_{k,p}(1), \tag{24}$$

and

$$F''_{k,p}(1)=\frac{2-kp-p}{(1-kp)^3}kp^2. \tag{25}$$

Substituting the equations (21),(22),(24) and (25) into (23), we have

$$\mathrm{var}(L) = \frac{pq}{(1-kp)^3}\left(\log p + (k-1)\log q\right)^2 .$$

Summarizing the previous results, we established the following theorem.

Theorem 3 The expectation $\overline{L}$ and variance $\mathrm{var}(L)$ of the ideal length of codewords for k-ary tree generated by the extending probability $0 \leq p \leq 1/k$ are given by

$$\overline{L} = \frac{h(p)}{1-kp}, \tag{26}$$

and

$$\mathrm{var}(L) = \frac{pq}{(1-kp)^3}\left(\log p + (k-1)\log q\right)^2 . \tag{27}$$

5 Case of Unary-Binary Tree

Next, we assume that the possible number of branches is zero, one and two with probabilities p_0, p_1 and p_2, respectively. we call the restricted tree as unary-binary tree. The number of unary-binary tree with n_1 nodes having one outgoing branch and n_2 nodes having two outgoing branches is given by:

$$c_{(1,2),(n_1,n_2)} = \frac{1}{n_1+2n_2+1}\binom{n_1+2n_2+1}{n_0,n_1,n_2}. \tag{28}$$

where $n_0 = n_2 + 1$ is the number of leaves. For this case, we introduce a generating function:

$$F_{(1,2),(p_1,p_2)}(x,y) = \sum_{n_1,n_2=0}^{\infty} c_{(1,2),(n_1,n_2)} p_0^{n_2+1} p_1^{n_1} p_2^{n_2} x^{n_1} y^{n_2}, \tag{29}$$

where $p_0 = 1 - p_1 - p_2$. Then, we have the recursion:

$$F(x,y) = p_0 + p_1 x F(x,y) + p_2 y F(x,y)^2, \tag{30}$$

by using the symbolic method. Solving the above equation (30), we get

$$F(x,y) = \frac{1 - p_1 x - \sqrt{(1-p_1 x)^2 - 4p_0 p_2 y}}{2p_2 y}. \tag{31}$$

Therefore, we can show the following theorem.

Theorem 4 The probability of the event E_f of having a finite unary-binary tree for the branching probability (p_0, p_1, p_2) is given by

$$\begin{aligned} \Pr\{\mathrm{E}_f\} &= F_{(1,2),(p_0,p_1,p_2)}(1,1) \\ &= \begin{cases} 1 & \text{for } 0 \le p_1 + 2p_2 \le 1 \\ \frac{p_0}{p_2} & \text{for } 1 \le p_1 + 2p_2 \end{cases} . \end{aligned} \tag{32}$$

From (30), we have

$$\begin{aligned} F_x(x,y) &= p_1 F(x,y) + p_1 x F_x(x,y) \\ &\quad + 2p_2 y F(x,y) F_x(x,y), \end{aligned} \tag{33}$$

$$\begin{aligned} F_y(x,y) &= p_1 x F_y(x,y) + p_2 x F(x,y)^2 \\ &\quad + 2p_2 y F(x,y) F_y(x,y), \end{aligned} \tag{34}$$

where F_x, F_y are the partial derivatives with respect to the variables x, y, respectively. From these functional equations, we can evaluate the expectations of the numbers N_1 of unary nodes and N_2 of binary nodes:

$$\begin{aligned} \mathrm{E}N_1 &= \sum_{n_1,n_2=0}^{\infty} n_1 c_{(1,2),(n_1,n_2)} p_0^{n_2+1} p_1^{n_1} p_2^{n_2} \\ &= F_x(1,1) \\ &= \frac{p_1}{1-p_1-2p_2}, \end{aligned} \tag{35}$$

and

$$\begin{aligned} \mathrm{E}N_2 &= \sum_{n_1,n_2=0}^{\infty} n_2 c_{(1,2),(n_1,n_2)} p_0^{n_2+1} p_1^{n_1} p_2^{n_2} \\ &= F_y(1,1) \\ &= \frac{p_2}{1-p_1-2p_2}. \end{aligned} \tag{36}$$

Collecting the above evaluations, we can establish the expected ideal codeword length $\overline{L}$ for the unary-binary trees that corresponds to the entropy of the tree.

Theorem 5 The expectation $\overline{L}$ of the ideal length of codewords for unary-binary tree generated by the branching probability $P = (p_0, p_1, p_2)$ is given by

$$\begin{aligned} \overline{L} &= \mathrm{E}\{-(N_2+1)\log p_0 - N_1 \log p_1 - N_2 \log p_2\} \\ &= \frac{H(P)}{1-p_1-2p_2}, \end{aligned} \tag{37}$$

where $p_0 = 1 - p_1 - p_2$ and $H(P) = -p_0 \log p_0 - p_1 \log p_1 - p_2 \log p_2$ is the entropy of branching probabilities P.

6 Entropy of i.i.d. Tree Source

More generally, we can establish the final result[6] for the i.i.d. tree source with respect to a distribution P.

Theorem 6 Assume that the number N of branches emitting from each node obeys the probability distribution $P = (p_0, p_1, \ldots)$. Under the condition $\mathrm{E}N \leq 1$, the expectation $\overline{L}$ of the ideal length of codewords for tree generated by the branching probability distribution P on non-negative integers is given by

$$\overline{L} = \frac{H(P)}{1 - \mathrm{E}N}, \tag{38}$$

where $H(P)$ is the entropy of branching probabilities P, and $\mathrm{E}N$ is the expectation of the number of branches.

Proof omitted. □

参考文献

[1] K.Kobayashi and T.S.Han, "On the Pre-order Coding for Complete k-ary Coding Trees," *Proceedings of 1996 International Symposium on Information Theory and Its Applications*, pp.302–303, 1996.

[2] K.Kobayashi, H.Morita and M.Hoshi, "Enumerative Coding for k-ary Trees," *Proceedings of the 1997 IEEE International Symposium on Information Theory*, p.423, 1997.

[3] K.Kobayashi, H.Morita and M.Hoshi, "Information Theoretic Aspects on Coding of Trees," *Proceedings of Memorial Workshop for the 50th Anniversary of the Shannon Theory, held at Yamanashi*, pp.43–45, 1999.

[4] K.Kobayashi, H.Morita and M.Hoshi, "Coding of Ordered Trees," *Proceedings of the 2000 IEEE International Symposium on Information Theory*, p.15, 2000.

[5] K.Kobayashi, H.Morita and M.Hoshi, "Percolation on k-ary Tree," *Abstracts of the Opening Conference on General Theory of Information Transfer and Combinatorics*, ZiF at Bielefeld, 2002.

[6] K.Kobayashi, H.Morita and M.Hoshi, "A Tree Source and its Entropy," *Proceedings of the 2003 IEEE International Symposium on Information Theory*, p.28, 2003.